Audio Quality Terminology

ABSTRACT

The terms described herein relate to audio quality artifacts. The intent of this document is to ensure Avaya customers, business partners and services teams engage in effective communication involving audio quality related issues.

©2005 Avaya Inc. All Rights Reserved.

1 Introduction

This document defines a variety of terms used to describe voice-related artifacts experienced in telephony. It is expected that this terminology will be used primarily by Avaya business partners and Avaya Global Services teams to facilitate the interpretation and understanding of voice-related problems experienced in the field.

2 Audio processing components and terminology

In a typical telephony call, speech from talker to listener often passes through the following processing components and in the order identified in Figure 1.

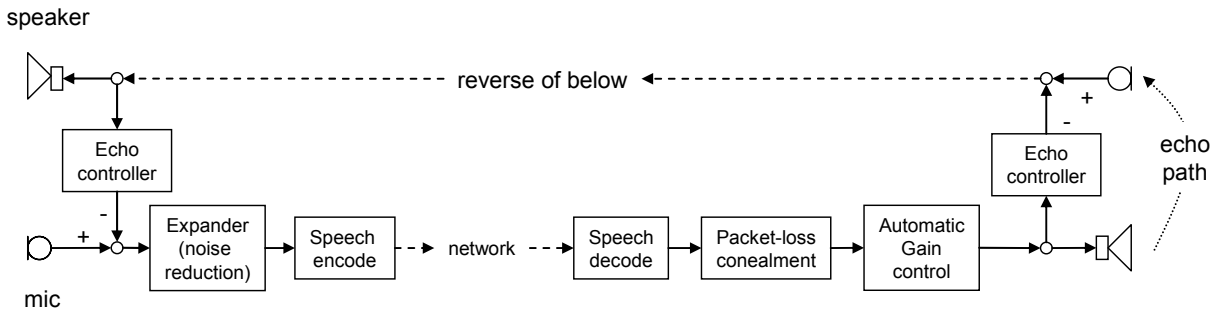

Figure 1. Components of the end-to-end speech path. The upper path is identical to the lower path, but reversed in order. The network could be TDM, packet (VoIP), or a combination of the two.

The talker's voice enters at the microphone on the left side of Figure 1, then to the microphone expander, voice coder, network transport, voice decoder, packet-loss concealment, echo controller, automatic gain control and, finally, the listener's ear.

2.1 Audio Processing Components

2.1.1 **Echo controller:** broad term meaning an echo canceler, echo suppressor, or a combination of the two. Speakerphone algorithms are also included. An echo controller prevents a talker from hearing distant reflections (echoes) of his/her own voice, reflections caused by acoustic or electrical reflection points within the telephone network and end-user equipment. Echo controllers are often only partially successful, and this is why echo is sometimes heard even though the call path is known to include echo controllers. Often, people use the term "echo canceler" when in fact what is being referred to is an echo controller.

2.1.2 **Echo canceller:** a software or hardware implementation of a digital signal processing algorithm designed to model and subtract-out – or cancel – the reflection, or echo, of a speech signal. Strictly speaking, an echo canceler does not introduce attenuation or suppression into the speech paths to reduce the loudness of echo. The term *canceler* refers to an adaptive digital filter that models the physical echo path and subtracts that (excited) model from the return speech path.

©2005 Avaya Inc. All Rights Reserved.

2.1.3 **Echo suppressor:** like echo canceler, above, except the echo level is reduced or eliminated by applying suppression or attenuation to the return speech channel. The use of attenuation causes other audio artifacts, including chopping or clipping of speech utterances and/or pumping of the loudness level of a caller's speech.

2.1.4 **Microphone expander, and or noise reduction:** a microphone expander is a traditional and relatively simple method of improving the speech-signal-to-background-noise ratio emanating from the microphone path. An expander attenuates weak room background noises while passing unaltered the relatively loud speech of the talker addressing the handset (or headset, or speakerphone).

2.1.5 **Speech coder (encoder and decoder):** the raw speech signal, once digitized, is often digitally encoded for transmission into the telephone network. Encoding has one purpose, namely, to reduce the bits-per-second rate of transmission required to communicate voice from one end to the other. Highly compressive codecs, such as a G.729 codec, reduce speech to a low transmission rate (8000 bits-per-second), but sacrifice voice quality in doing so. Higher voice quality is experienced in systems using the traditional G.711 codec (mu-law codec), since G.711's higher transmission rate of 64,000 bits-per-second better captures the nuances of speech. Regardless of coding scheme, at the receiving side, the speech decoder reconstructs (an approximation to) the original speech for playback.

2.1.6 **Packet-loss concealment:** often combined with speech decoders. When the network path includes packet-speech transmission links, like VoIP, speech packets can be lost because of network failures. In such cases, a concealment algorithm attempts to fill-in missing speech samples. Concealment can work well when the rate of lost speech is very low, say, less than 2% of transmissions.

2.1.7 **Automatic gain control:** automatic gain control devices apply signal gain or loss automatically in an attempt to keep the speech sound level at the listener's ear relatively constant. Therefore, AGC boosts low-level speech while reducing speech levels that are too loud. Such devices have been used for decades in audio broadcasting and recording applications.

3 Terminology for voice-related artifacts

3.1 Speech distortions

3.1.1 **Distorted speech:** speech accompanied by an unnatural buzzing or raspy sound. A classic example of distortion occurs in the case of a far party who is speaking too loudly or too close to the handset or headset microphone. The far party's speech saturates either the mechanical or electrical capabilities of the handset, causing overload distortion or amplitude clipping

3.1.2 **Muffled speech:** speech that has an unnatural loss of high-frequency content. Muffled speech may be caused by, for example, poorly designed microphone assemblies in handsets (in particular, wireless handsets) and low-bit-rate speech coders.

3.1.3 **Reverberant speech (also hollowness or speaking-in-a-tunnel effect):** sounds like the person speaking is in a barrel or large empty room. This can be the case when the talker is using a speakerphone, but it can also be the case when there is network echo, e.g., in a teleconference without echo control.

©2005 Avaya Inc. All Rights Reserved.

3.1.4 **Synthetic, Mechanical, or Robotic Voice**: this can be very subtle or very severe, or very consistent or intermittent. In the most severe case, the pitch information has been lost making the speech sound monotonic and robotic. Recognizing who is speaking is often difficult.

3.1.5 **Amplitude clipping:** see definition for *distortion* above**.**

3.1.6 **Clipping:** portions of the speech signal are not heard. This can occur in packet-switched networks when, for example, large numbers of successive speech packets are not received because of excessive network congestion. Common in wireless phones, where the RF-signal strength fades as the user moves within the environment.

3.1.7 **Clipping during double-talk:** clipping, as defined above, but heard only when both parties of a telephone call talk at the same time. When it occurs, this effect is almost always caused by the excessive use of echo suppression (see definition) at some point within the network. In this case, clipping of speech utterances is not caused by lost speech packets or, in the case of wireless phones, RF fades, though those artifacts may also be present in the same call.

3.1.8 **Stutter:** this is often used to describe an effect caused by repetition of short bursts of noise or speech, such as "da-da-da-da…" or "fa-fa-fa-fa…" Stutter distortion can occur in packet-speech networks when one or more network elements (e.g., router or switch) become a bottleneck to the timely transmission of speech packets.

3.1.9 **Speech-level pumping:** pumping is often used to describe a varying speech-loudness level, that is, were the speech gets louder, softer, then louder again, etc., over the course of a call, often over a period of just several seconds. Automatic gain control devices can cause audible and distracting pumping.

3.2 Noise and Other Phenomena

3.2.1 **Hiss or white noise:** relatively natural-sounding noise containing energy at all frequencies. Low-level, idle-channel hiss noise can be perceived on nearly every telephone call when no person is speaking.

3.2.2 **Static:** impulsive, ticking noise, similar to the sound of an AM radio when tuned to a very weak or nonexistent radio station. In a packet-speech network, can be caused by lost speech packets and/or bit errors. May also be used to describe power-line hum (see definition below).

3.2.3 **Motor boating:** repetitive noise that is separate and distinct from the talker's voice. Motor-boat noise differs from static in that it is repetitive or non-random.

3.2.4 **Hum:** sounds like humming, as in "Hmmmmm…" Hum noise often occurs when a source of 50 Hz or 60 Hz electrical power is located near a telephone. The power source emits an RF (radio frequency) field that induces a hum-like noise that is heard through the phone's handset/headset earpiece or speakerphone loudspeaker.

3.2.5 **Distorted Music-on-Hold or Dialtone:** low-bit-rate codecs such as G.729 , and G.723 , were created to efficiently encode and transport speech but not music (or other non-speech signal such as tones). Thus the usage of these and other codecs may distort

©2005 Avaya Inc. All Rights Reserved.

and ruin the music signal or non-speech signal. This can be subtle or severe depending on the music source.

3.3 Echo

There are only two physical sources of echo in telephony: electrical echo (or network echo), and acoustic echo. Electrical echo is caused by a reflection of the speech signal at 2-to-4-wire hybrid circuitry. This circuitry is present in analog trunk cards, and it also exists deep within the PSTN (at customer premises, for example). Acoustic echo is caused by the physical coupling (air path, appliance-body path) between a loudspeaker and a microphone, for example, in a speakerphone, a handset and a headset. Whether or not a talker actually perceives electrical or acoustic echo depends on the loudness of his/her reflected voice signal and the roundtrip delay that that reflection suffers. The loudness of the reflection at the point of reflection depends upon the electrical impedance mismatch, for electrical echoes, and the acoustic gain of the loudspeaker-to-microphone path, for acoustic echoes. The roundtrip delay is a function of the path the reflected signal traverses, which in turn is a function of the call topology.

3.3.1 Electrical echo, also called network echo: reflection of a talker's speech signal at a point of 2-to-4-wire conversion caused by an impedance mismatch at the point of analog-to-digital conversion.

3.3.2 Acoustic echo: reflection of a talker's speech signal at an acoustic endpoint caused by the acoustic coupling between the loudspeaker and microphone.

3.3.3 Constant echo: when talking, the perception of echo with every utterance. Such cases occur when there is a physical electrical or acoustic echo path but no echo controller in the call topology to control echo. Additionally, constant echo may result even though an echo controller is known to be in the call path; this indicates a complete failure of the echo controller, usually because the capabilities of the echo controller are exceeded (e.g., the echo tail length exceeds the specifications of the echo controller).

3.3.4 Intermittent echo: when talking, the occasional perception of echo. Intermittent echo often caused by the intermittent failure of an echo controller in the call path. The echo suppressor within the echo controller may fail to engage (to apply echo attenuation) when necessary, with the result that short bursts of echo become audible. In acoustic echo control applications (speakerphone) in which people or objects close to the speakerphone are moving, the change to the physical echo path often results in audible intermittent acoustic echo to listeners at the other end of the call.

3.3.5 Residual echo: when talking, the perception of very low-level (quiet) echo. The echo could be either constant or intermittent. Residual echo can be caused by PSTN electrical echo that is not entirely removed by the echo controller in the call path.

3.3.6 Distorted or buzz-like echo: when talking the perception of a distorted echo or buzz-like sound. This can be caused by a non-linear echo source. An example of this is saturation distortion at an analog trunk interface. In this case, signals low in amplitude are reflected cleanly, but signals high in amplitude are returned with significant distortion making it difficult for an echo canceler to control echo. Such distorted echo can be perceived constantly or intermittently, depending on the degree of distortion and the echo canceler(s) involved.

3.3.7 Slapback or kickback acoustic echo: this is strictly a phenomenon of acoustic echo. With speakerphones, slapback or kickback echo is the intermittent echo perceived at the ends of one's utterances. This can occur with both older-model half-duplex

©2005 Avaya Inc. All Rights Reserved.

speakerphones and newer-model acoustic-echo canceling speakerphones. For example, a talker speaking into a handset utters the phrase “Please send me the check” and perceives echo primarily at the end of his/her sentence. This echo is described as hearing just the sound “eck” or “k” of the word “check,” or as a slapping sound such as that made by slapping one’s palm against a desktop. Commonly, slapback/kickback echo is caused by acoustically reverberant rooms. Large offices and conference rooms can have long reverberation times. In such rooms, the speakerphone senses at its microphone a reverberated version of the word “check” (our prior example) several tens or even hundreds of milliseconds after the far talker has finished saying the word “check.” The speakerphone algorithm detects this reverberated speech at its microphone, detects no speech at its receive-path driving the loudspeaker, and decides to transition to transmit mode. The reverberated version of "check" is transmitted back to the far talker, where it is perceived as echo.

3.3.8 **Sidetone:** in handsets and headsets, a portion of the microphone energy is fed back to the earpiece so that the user of the handset/headset does has a psychoacoustic experience that simulates the case in which the user's ear is not occluded by an object (the handset earpiece). Without sidetone injection, the user experiences the psychoacoustically bothersome condition that can be demonstrated to oneself by pressing a finger into one ear while speaking. With one ear occluded, the sound of one’s own voice is dominated by the path through the interior of the head (skull, etc.) instead of around the head, an effect that most people find objectionable.

3.3.9 **Hot sidetone:** in a handset or headset, microphone-to-earpiece sidetone injection is not normally noticed. Some digital phones, in particular, IP phones in which the internal audio processing frame rate is 5 ms or greater, inject sidetone with an appreciable delay (e.g., 5 ms) in the microphone-to-earpiece signal path. This delay causes the sidetone to sound reverberant and/or louder than normal, or hot. Though hot sidetone is a type of echo source – because some people may use the term “echo” to describe hot sidetone – it is generated local to the telephone, not at some point within the telephone network.

3.3.10 **Short-path acoustic echo, short-path electrical echo:** acoustic or electrical echo that occurs in a very short roundtrip call topology. This type of echo is commonly described as a hollow sound or sound of speaking in a barrel (see 2.1.3). In a digital-to-digital phone call (think DCP-to-DCP), station-to-station, the roundtrip delay is usually very small, less than 10 ms. Some digital speakerphones produce significant acoustic echo, which is not canceled, suppressed, or otherwise controlled in this simple call topology. In these cases, and depending on the volume setting of the far-party’s speakerphone and near-party’s listening handset, the near party may perceive echo and refer to this as hot sidetone. Again, this is truly acoustic echo from the speakerphone but is returned to the talker with such a short roundtrip delay that it is perceived as hollowness or reverberance rather than as classic echo. Because of the short roundtrip delay in this case, it can be difficult to distinguish between *hot sidetone* (see definition) and short-path echo.

©2005 Avaya Inc. All Rights Reserved.